Data Mining

BY HEIDI KUZMA AND SHEILA VAIDYA
DATA MINING

BY HEIDI KUZMA AND SHEILA VAIDYA

Each day, the facilities making up the Comprehensive Nuclear-Test-Ban Treaty’s (CTBT) International Monitoring System (IMS) collect approximately 10 gigabytes of seismic, hydroacoustic, infrasound and radionuclide data, which are relayed via a secured satellite network to the International Data Centre (IDC) in Vienna, Austria. An automatic processing system sifts through the data, flagging all potential events which could be related to a suspected nuclear explosion, reporting them in a series of Standard Event Lists: SEL1, SEL2 and SEL3. The events in the final list, i.e. SEL3, must be carefully examined by a team of highly trained professional analysts before they are published in a Reviewed Event Bulletin (REB). In the process of putting together the REB, analysts discard roughly half of the SEL3 events and make many more corrections. The data, lists and bulletins are stored in a database at the IMS, which is large enough to archive 10 years worth of data. As technology evolves, the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) mandates the incorporation of new techniques into IDC data processing to continually improve the quality of its products.

Computing techniques have improved dramatically over the recent past and are constantly evolving. A variety of algorithms have been developed for data mining, a catch-all phrase for the use of automated algorithms such as clustering, classification, and statistical inference to find patterns and relationships in large amounts of data. Data fusion algorithms synergistically combine different types of related data so conclusions drawn from a large data set can be greater than sums of its parts. Data management systems intelligently store, retrieve and distribute processing in order to facilitate access to and interpretation of enormous amounts of data.

During preliminary International Scientific Studies (ISS) meetings in 2008, it was recognized that up-to-date algorithms for data mining, data fusion and data management could be used to improve many aspects of the data processing pipeline at the IDC. In addition, it was established that these algorithms could help in furthering the IDC’s mission to make available its data, infrastructure and global sensing environment for other scientific, humanitarian and security related efforts. A direct dialogue was opened between the IDC and computer scientists. As a result, many posters were presented at this year’s ISS Conference (ISS09). Topic coordinators Dr. Sheila Vaidya from Lawrence Livermore National Laboratories and Professor Arno Siebes from the University of Utrecht reflected on the role of Data Mining in two separate panel sessions and Professor Stuart Russell, chair of the Computer Science Department at the University of California, Berkeley, gave a well received keynote speech.

DATA MINING: GENERAL THEMES

The dominant theme of the Data Mining posters was to recognize that analysts are learning patterns in data that the automatic system at the IDC currently ignores. In an introductory poster, Pearce et al. (DM-12/A) pointed out that analysts build up empirical knowledge of specific source-station paths. This allows them to better recognize meaningful signals, discard events and make use of negative data, that is, the absence of expected signals at particular distances or types of stations. Selby and Bowers (DM-10/A) showed that the seismic station ARCES in Norway did not detect the 2006 Democratic People’s Republic of Korea (DPRK) explosion because station-specific correlated noise masked the signal to the automatic processing system, but not to a trained human analyst who knew where to look for it.

Much of the work presented at ISS09 involved training automatic algorithms to find patterns in the SELs in order to better classify events as “true” (i.e. the event will be retained by the analyst and finally reported in the REB) or “false” (the event will be discarded). Figure 1 illustrates an overlay of events from 2008 that were discarded by the analysts.
Carluccio et al. (SP-12/C) plotted features of events in the SEL1, such as latitude, longitude and error ellipse, against similar features in the REB in a poster which won the overall Best Poster Award. By examining patterns of true (green) and false (red) events in a set of graphs, they motivated the training of a neural network to distinguish between events. Other off-the-shelf, quickly implemented modern classification methods also easily discriminated between true and false events (Kleiner et al., DM-02/A; Schneider, DM-05/A; Procopio et al., DM-07/A; Gaillard et al., DM-14/A). The authors of these posters tested a number of state-of-the-art classification algorithms including Support Vector Machines (SVMs), Naive Bayes Classifiers and Decision Trees, to achieve correct classification rates on the order of 80 percent to 85 percent. Kleiner et al. pointed out that obtaining an 84 percent correct classification rate could potentially reduce the current analyst culling effort by as much as 57 percent.

Analysts currently spend a considerable amount of time evaluating, re-labeling, and re-associating seismic detections listed in the SEL3. If these phases were more correctly listed, the automatic event associations and origins would also be more accurate. Using the same techniques as they did for false event detection, Kleiner et al. and Schneider showed that algorithms can be trained to automatically test SEL3 phase labels with over 90 percent correct classification accuracy. Kuzma et al. (DM-03/A) trained an SVM for regression instead of classification in order to learn
regional effects on travel times and improve phase identification and SEL3 event classification. Ohrnberger et al. (DM-04/A) designed an algorithm to detect and accurately label the phase of arrivals using features derived from a continuous wavelet transform of the raw waveform data. Because seismic phases tend to occur in a precise order, they borrowed techniques from the automatic speech recognition community to build “grammar” constraints into their model.

The final product from the IDC is a Screened Events Bulletin (SEB), which does not include events from the REB that are likely of natural origins (earthquakes). Tuma and Igel (DM-06/A) used SVMs to discriminate between natural events and explosions in hydroacoustic data. Materni et al. (SEISMO-22/J) applied discriminant analysis to low-frequency regional seismic surface waves in order to distinguish between natural vs. man-made events. Horin and Steinberg (SP-06/C) tried to find patterns in the relationships between SEL3 and REB which could be identified as the work of individual analysts.

Several purely scientific posters mined IDC data sets to further understand the Earth and its inhabitants. It was shown that patterns in hydroacoustic data can be used to identify the songs of whales and follow their seasonal migrations (Flore et al., HYDRO-17/H). Harris et al. (HYDRO-12/H) tracked the calls of Antarctic, pygmy and Sri Lankan blue whales.

..........................................................

STATISTICAL INFERENCE

Instead of simply classifying events, a second set of posters used techniques of statistical inference to mimic an analyst’s intuitive grasp of what constitutes a “high probability event,” i.e. a pattern of detections that is probably a true event given the measurement characteristics of the IMS network and physical properties of the Earth. The basic method of statistical inference (also known as Bayesian inference) is to compute the posterior probability of an event taking into account all the variables that contribute to the evidence for that event. The output of a Bayesian inference algorithm is a probability distribution which can be interpreted as a set of likely outcomes. Assembling a probabilistic model is not as straightforward as applying a general purpose classifier, but it holds the promise of being able to interpret data in a way that incorporates experience, physics and eventually, multiple types of data.

In a set of posters which also formed the basis for the keynote speech on data mining by Stuart Russell, Arora et al. (DM-08/A and DM-09/A) used Bayesian inference to find patterns of detections in the SEL that, with high probability, are consistent with seismic events of given locations and magnitudes, explaining that the inference process works by continually postulating events, and then creating, modifying, and deleting these events using an approach known as Markov Chain Monte Carlo (MCMC) sampling. As reproduced in Figure 2, peaks in the posterior or probability density correspond to events that were confirmed by analysts in the REB. False SEL3 events, which had to be discarded by the analysts, were not generated by the inference engine.

Myers and Johannesson (DM-11/A) presented a software package called Bayesloc which uses a similar technique to locate multiple events from a stream of automatic detections. By studying the posterior distribution of various factors pertinent to event location, they were able to decompose measurement error into various components so that data from stations with high quality detections could receive more weight, producing events with half of the error of conventional processing.

Also using Bayesian methods, Schorlemmer (SEISMO-42/K) derived the probability distributions of events at each station in the IMS so as to produce probability maps for the detectability of earthquakes. Liszka (DM-15/A) created a probabilistic propagation model to improve the interpretation of infrasonic observations. Carrigan and Johannesson (OSI-10/B) used posterior probability as a means of narrowing the search area for Surface Ground Zero during an On-Site Inspection.

..........................................................

DATA FUSION

Despite general consensus that the evidence from all IMS sensors should be used in concert, seismic, hydroacoustic, infrasonic and radio-nuclide data are currently processed indepen-
dently, making data fusion a manual task and consequently a fertile area for research. A few authors addressed the need for data fusion directly. Jepsen and Fisk (HYDRO-06/H) recognized that a lack of high-frequency hydroacoustic energy indicated that the seismic events were more likely to be natural earthquakes than nuclear explosions. Maceira and Ammon (SEISMO-52/K) fused gravity and seismic data, and, in an unusual poster, Bossu et al. (SEISMO-28/K) correlated the IP addresses of traffic on the website of the European Mediterranean Seismological Centre to the geographical areas in which an earthquake was felt.

**DATA MANAGEMENT**

A consistent thread which permeated many of the informal discussions at ISS09 was the need for faster and more comprehensive access to the IDC database. Simpson and Willermann (SEISMO-56/K) detailed the Incorporated Research Institutions for Seismology (IRIS) consortium Data Management Center as an example of a seismic data centre providing free and open access to many global networks through a common set of tools. Pesaresi et al. (DM-13/A) presented Antelope, a commercial software suite which can be used for real time data exchange.

An entirely new paradigm for the storage and exploitation of seismic data was suggested by Liu (DM-16/A). Currently, seismic records can only be retrieved from the relational tables in the IDC database using a fixed set of attributes such as station name, channel number, latitude, longitude etc. It is not possible, for instance, to ask the database to retrieve all waveforms that correlated well to a particular set of arrivals from a particular event. In order to ask this type of question, the entire database must be downloaded to a researcher’s own computer where the correlations must be computed locally. Liu proposes developing a distributed database using a framework similar to ones used commercially (by Google in particular), which take advantage of parallel processing and multiple nodes to return records based on complicated and

---

**Figure 2**

Bayesian inference can be used to provide a posterior probability distribution giving the probability that events occurred given features in the IDC data. Using this technique, events show up as regions of high probability (in red on the large map). In Arora et al., (DM-08/A), two hours of SEL3 data were analyzed. In this example, reproduced from the poster, the algorithm is able to postulate new events (yellow stars) that were missed by the automatic processing but were confirmed by the REB.
flexible queries and which are easily scalable to accommodate large amounts of data in different formats.

MOVING FORWARD

Zhongliang and Suhadolc (SEISMO-32/K) liken the role of the IMS to that of the Hubble Space Telescope: it provides a well-functioning global observation facility with the potential to make fundamental contributions to basic research. At present seismology is in a period of fast development, which they credit to several factors among which are the “continuous accumulation of high-quality data and the application of new technologies in observation and data analysis.” By extension, the effort which is now clearly underway to apply leading-edge technologies for data mining, fusion and management at the IDC promises not only to significantly improve its automatic processing system, but also to lead to a new and better understanding of the Earth. In addition to helping fulfill the present mission of the CTBTO, this effort will advance science and serve humanity worldwide.

Many fruitful areas for research were identified at ISS09. These include but are not limited to:

- The establishment of the best features to use for automatic event classification, including the generation of new features;
- The refinement of probabilistic inference algorithms to take many different kinds of data into account;
- The application of data mining to the CTBT’s On-Site Inspection regime;
- The use of data mining to learn about effective regional corrections for seismic data;
- The development of systems with the capability of analyzing full-waveform data;
- The design of fast, effective databases for the storage and distribution of IDC data;
- The establishment of a centre to champion, coordinate and support advanced research and data analysis.

Such a virtual Data Exploitation Centre (vDEC), proposed by Vaidya et al. (DM-01/A), would create an incubating environment for ISS research. A vDEC would also provide a development clearinghouse for software and hardware infrastructure in support of the IDC and OSI missions, seeking contributions from participating Member States as well as forming strategic partnerships with national agencies with synergistic agendas.